Considerations of Methodology and Human Factors in Rating a Suite of Translated Sentences

Leslie Barrett
Transclick, Inc.
535 W. 34th st. New York, NY 10001
leslie@transclick.com

Abstract

This paper describes and analyzes the results of rating a suite of test-sentences for an Arabic-English/English-Arabic translation system at Transclick, Inc. The task of rating this suite presents a challenge in that it is composed entirely of sentences that are unrelated to one another, and thus certain typical evaluation methods do not easily apply. The suite is rated with a view to exploring evaluation methods for this particular type of data, and observing human qualitative judgments of the data, rather than rating the actual quality of the MT system used. In particular, this paper discusses the degree of inter-tester agreement, and compares our findings to those of other studies where inter-tester agreement on language tasks has been analyzed. We suggest some possible reasons for the relatively low agreement values, and propose future strategies to address the problem.

1.0 The Task and Test Suite

Judging the output quality of any machine translation (MT) system requires addressing the grammatical and lexical properties of the domain, creating a test-suite covering those properties and finding a reliable metric of quality. This paper reports on the partial result of an ongoing evaluation of a commercial off-the-shelf MT product for use in an English-Arabic/Arabic-English speech-to-speech translation prototype[1]. In this case, we developed a preliminary test suite for the military command-talk domain containing 28 English sentences of three sentential types. The purpose of this very small preliminary suite was to get a sense of the system's ability to handle certain sentence types, and to find the degree of tester agreement on scoring, rather than to provide a qualitative evaluation of the system itself. For this purpose, a larger corpus of sentences is being compiled for a second evaluation. Overall, the results from the preliminary study were intended to help us judge our own evaluation methods, and test-suite design principles. The three sentential types were selected in the percentages that were likely to be representative of the domain. We chose sentences from a training corpus provided by the U.S. Army Research Laboratory. Table 1 shows the distribution of the suite by sentential type.

Table 1.
Distribution of Sentential Types in the Test Suite

Sent Type	No. of Sentences	% of Corpus	Mean Length
Imperative	7	0.25	5.3
Interrogative	8	0.29	6.8
Declarative	13	0.46	9.5
Overall	28	1	7.2

[1] Transclick, Inc. is engaged in the development of a bi-directional English/Arabic speech-to-speech translation system prototype.

Apptek, Inc provided our machine-translation component, for both English-Arabic and Arabic-English translation.

Three bilingual testers rated each sentence in the test-set on a four-point scale. On this scale, a rating of "1" was a perfect translation, "2" was a slightly flawed translation, "3" was a heavily flawed translation and "4" was an incomprehensible message (or completely divergent from source). Similar types of 4-point subjective evaluations have been used previously in Sumita et al. 1999, and on a corpus of sentences more recently in Akiba et al. 2001. Although our scale approximates the criterion of *clarity* discussed in Miller and Vanni (2001), our scoring differs somewhat from this metric in that fidelity to the source sentence is considered here. Our testers were advised that grammatical "flaws" were meant to include deviance in any three metrics of correctness (i.e. lexical, syntactic and semantic[2]) modeled on those outlined in Nyberg, Mitamura and Carbonell (1994). We did not test separately, therefore, for syntactic or morphological quality or measures of coherence including Rhetorical Structure Theoretical metrics (Vanni and Miller 2001). We found such criteria, although successful for other MT-evaluation studies, are more appropriate to translations of data chunks much larger than sentences. ere less concerned in this preliminary test-suite with counting grammatical flaws for the purposes of rating the system than we were with comparing human judgments on just one (subjective) parameter. Each sentence in our test-set received one rating from each tester in each language direction. The test suite was created in English. For the purpose of rating English translations, one of the testers[3] translated the suite into Arabic. Both suites were run through the Apptek Transphere® system, and the outputs were evaluated by the testers.

2.0 Sentence Scores

Table 2. shows the testers' scores by sentential types in the English to Arabic direction. As one can observe by comparing the results in Table 1 with those of Table 2, mean sentence length correlates positively with mean score within sentential categories.

Table 2.
Mean Score by Sentence Type: English-Arabic

Sent Type	Mean Score	Mean Length
Imperative	1.76	5.3
Interrogative	1.95	6.8
Declarative	2.65	9.5

[2] Semantic deviance in this study includes a meaning that diverges from the source, even if grammatically sound.

[3] This tester did not evaluate the Arabic-English test set.

Using just length as a factor, however, sentences which are 9 or more words in length scored highest (i.e. indicating the poorest translation score), but length alone did not correlate exactly with score. Results are shown in Table 3.

Table 3.
Mean Score by Sentence Length: English

Sent. Length	Mean Score
2-5 wds	1.95
6-8 wds	1.75
9+ wds	3.05

In the Arabic to English direction, where Arabic sentences do not match English in length, the results for scores by sentential category still tended to be similar. Table 4 below shows these results:

Table 4.
Mean Score by Sentence Type: Arabic-English

Sent Type	Mean Score	Mean Length
Imperative	1.5	4.7
Question	1.75	5.0
Declarative	2.34	6.5

In the Arabic-English direction, however, length alone did correlate well with score, as shown in Table 5.

Table 5.
Mean Score By Sentence Length: Arabic

Sent Length	Mean Score
2-5wds	1.63
6-8wds	1.95
9+ wds	3.67

There was no association found between running order and sentence score in any of the testers' ratings in either language-direction.

The small size of the corpus caused category overlap between sentence type and length, and did not allow us to draw any meaningful information about the Apptek product's performance on different sentence types. The tendencies for the sentential types that emerged from this data set, looking at both language-directions, were that Imperatives score slightly better than other sentence types, while Declaratives score poorly. Based on these results, we concluded that the larger test suite going forward will need to balance the sentence-type inventory accordingly so that length and sentential type can be evaluated independently.

2.1 Tester Differences in Scoring

There was considerable variation between the testers' scores in the sample. Since we used three testers, we defined two types of agreement thereby, "full" and "partial". In a "partial" agreement, two testers have the same score, and for "full" agreement, all three testers have the same score. The expected frequencies for "full" and "partial" agreement, calculated by the same method in both language-directions, were .02 and .14 respectively. We called expected partial agreement the chance of 2 matches of 4 items in 3 trials[4]. Expected "full" agreement requires 3 matches in 3 trials. Thus, we used the following binomial distribution equation in (1) for obtaining r successes in N trials:

$$P(r) = \frac{N!}{r!(N-r)!}\pi^{r}(1-\pi)^{N-r} \quad (1)$$

The observed frequency for full agreement for Arabic-English was .21, and the observed frequency for partial agreement was .82. The observed frequency for full agreement in English-Arabic, was .21, and was .64 for partial. These figures are shown in Table 6:

[4] In our calculation the probability of a "Score" (1 through 4) was .25.

We were primarily interested in comparing the agreement values here to those in other types of studies. For example, the inter-annotator agreement in the SENSEVAL study (Veronis 1998) showed full inter-annotator agreements[5] of >40% on a word-sense task. A comparison of our results to these and others will be discussed in more detail in section 3.0.

Table 6
Agreement Values: All

	exp(full)	obs(full)	exp(p)	obs(p)
Eng-Ar	0.02	0.21	0.14	0.60
Ar-Eng	0.02	0.21	0.14	0.82

In an attempt to judge the reliability of our ratings, we measured the covariance between testers in both directions, and, where there were three testers (English-Arabic), looked at differences for each sentence score. The mean covariance was .58 between testers in the Arabic-English direction, and .66 between testers in the English-Arabic direction.

2.1.1 Defining "Score"

For the purposes of creating an expanded test suite suited to a meaningful evaluation of an MT engine, we needed a method of evaluating "score". It is not immediately obvious that this should be the mean score of *n*-testers, nor is it necessarily mode or median (see Akiba et al. 2001). We did consider "score" to be a range around the mean score, but chose a different method of calculating standard deviation[6]. We did not use the usual standard deviation measure because the spread reflected by this measure isn't always informative about the testers' judgments. Sentences with partial agreement show a much narrower spread around the mean by this method. Sentences with no agreement show the highest spread by both methods, but our method keeps the range tighter, falling within 1.0 to 4.0. This generally gives a more informative picture of the "true" translation score[7]. To address why our testers show the amount of agreement that they do, or whether the amount of agreement could be increased, however, we look to other similar studies for comparison. The next section will discuss the findings of other studies where human testers are compared on language-related tasks.

3.0 Agreement Between Testers in Other Tasks

We note that the amount of agreement between testers in this translation-evaluation task fell below the amount of agreement found between testers in other studies using language data.

Carletta (1996) argues that traditional methods of determining agreement between testers in various NLP tasks are not particularly effective. Some studies cited there (Passonneau and Litman 1993, Kowtko et al. 1992) compare testers' scores against an "expert" or "majority" opinion rather than comparing observed and expected agreements. Carletta (1996) uses the Kappa Statistic, previously used in content analysis, as an alternative to using comparisons with subjective standards to determine scoring reliability. The Kappa Statistic compares observed and expected values in the following formula in (3):

[5] In this case there were six testers, and "full" agreement is considered the same score by all six.

[6] We calculated the deviation from the mean of each score for each sentence in the English-Arabic direction according to this formula:

μS +/- $(1/\alpha^2)$
S=sentence score α= agree score
μ= mean

We established an "agree score" of "1" for *no* agreement, "2" for *partial* and "3" for *full* agreement. This effectively makes the spread for non-agreeing samples larger, and agreeing samples smaller.

[7] For example, sentence #3 in the test-set received a "2" from tester 1, a "1" from tester 2, and a "2" from tester 3. The mean was 1.67. The spread by our method was .50. The spread by the standard deviation method was 1.15. For partial agreement, our method reflects a tighter range around the mean, "crediting", in effect, the means of sentences where testers agree.

$$K = \frac{P(A)-P(E)}{1-P(E)} \qquad (3)$$

This coefficient measures pair wise agreement among a set of testers making category judgments, correcting for expected chance agreement, where P(A) is the proportion of times that testers agree, and P(E) is the proportion of times that agreement is expected. If there is no agreement other than chance, Kappa will be equal to 0, whereas if there is perfect agreement, Kappa will be equal to 1.

Although the studies cited by Carletta are using testers to determine discourse boundaries or prosodic phrase boundaries within one language, the issue extends easily to the present study. Using the expected and observed agreement values shown in the previous section, we found **K**=.18 for English-Arabic for "full" agree. For partial agree, we found **K**= .36. For Arabic-English we found **K**=.17 and **K**=.54 for full and partial agreement respectively. These figures, however, are less meaningful if not compared to other studies as well as other comparison methods. For example, using the Pearson's *r* test[8], in the Arabic to English direction we get a mean $r = .55$, in the English to Arabic direction, $r = .57$. Because this test does not take into account the expected agreement or exact score matches, it gives higher values for the data in the present study.

Veronis (1998) reported the results of testers' judgments on a word-sense-disambiguation task. The task involved six testers (annotators), and 600 words divided into 3 part-of-speech groups. Possible judgments, or scores, included rating the word as having one sense, multiple senses, or "don't know". The Veronis study recorded kappa values of between .37 and .67 depending on the part of speech. These are considerably higher than those of the present study. The Veronis study, however, does differ with the present study in certain non-trivial ways. For example "full" agreement was for all 6 judges having matching scores. The 6 testers had 3 possibilities to choose as a "score"[9](i.e., one sense, more than one and "don't know") not all the possibilities that would have occurred if the actual senses were counted. Taken either way, the expected value is lower than in the present study. Because of the high number of single-sense words in some categories however, (the study notes that adjectives in particular tend to have one sense) all choices are not necessarily equal.

In Brants (2000), inter-annotator judgments were recorded on a part-of-speech and structural annotation task. Six annotators tagged NEGRA, a German newspaper corpus with a tag set consisting of 54 part-of-speech tags, 25 grammatical phrase tags, and 45 grammatical function tags. Not all annotators were assigned the same sentences. Annotators were trained in advance to become familiar with the tag set.

For part-of-speech tagging, Brants used a measure of accuracy dividing the number of tokens tagged identically by the number of tokens in the corpus. The accuracy of this task by this measure was 98.57%. This agreement rate is high, but not uncommon for POS-tagging tasks. A study by Voutilainen (1999) showed similarly high rates for inter-annotator agreement. Brants notes differences in agreement values based upon the identity of the tag. In particular, he notes that the highest rates of annotator differences occur with tags that are infrequent in the corpus. With phrasal tagging, which yielded lower F-scores[10] and lower agreement generally as a task, the phrases with the highest rates of inter-annotator differences were also those that
tended to be least frequent[11].

[8] $$r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{(\Sigma X^2 - \frac{(\Sigma X)^2}{N})\ (\Sigma Y^2 - \frac{(\Sigma Y)^2}{N})}}$$

[9] According to the study, there were 3 POS categories, which could be alone or in any combination, plus the option of "don't know" (N, V, A, NV, NA, VA, NVA, 0), making the base probability .126. Therefore, the chance two testers of the six arriving at the same score would have been .016. However, "score" as defined in the Veronis study was agreement on *whether* or *not* a word had more than one sense.

[10] Brants notes that an analysis of tags causing high disagreement rates reveals that categories causing a high *absolute* number of differences do not coincide with categories causing high relative numbers of differences (high F-scores). The F-score is the harmonic mean of tester agreements.

[11] The author does not mention possible dependencies between tag scores and phrasal scores.

Comparing the raw agreement values for Arabic-English translation, our mean agreement score for full agreement was .21 for both language-directions. For English-Arabic, mean pairwise agreement[12] between the testers was .285 and .34 for Arabic-English.

4.0 Results of Comparison and Conclusion

Our results come closer in general to the agreement results of the Veronis (1998) study than to that of the Brants (2000) study, although our agreement values in general were below both. We suspect that, going forward, agreement scores for this kind of task will pattern more closely with scores in sense-tagging rather than POS-tagging tasks. Table 7 below shows a comparison.

Table 7.Comparative Results of Tester Agreement Based upon Test-Type[13]

	Brants	Veronis	Transclick
Kappa	NA	0.49	(0.18)/0.45
Pearson's	NA	NA	0.60
full agreemnt	0.98	0.45	0.21

Although we cannot draw any strong conclusions about tester agreement as a task based upon these results, we can make some meaningful observations that should provide direction for future work. We consider the scoring system, and tester's understanding of it primarily responsible for our results in this test. Both the Veronis and the Brants study had clearly defined scoring systems. In our study, we noticed that the polar values (1 and 4) represented 84% of the matches, while other matches represented only 16%. Therefore, it is likely that testers had a clear sense of when the translation of a sentence was unacceptably bad on the one hand, or perfect on the other hand. The middle ratings, however, representing either sentences that are "slightly" flawed (i.e. "2"), or sentences that are "heavily" flawed (i.e. "3"), were probably less clearly defined. It is possible that a clearer definition of the rating system would improve the agreement values in this kind of test in the future.

Other approaches to scoring, such as the correspondence model discussed in Ahrenberg and Merkel (2000), offer a less subjective rating system. The correspondence model is aimed at describing and quantifying structural and semantic relations between source text and translation. The application of this model in Ahrenberg and Merkel (2000), however, is sufficiently complex that testers would require extensive training and need some background in linguistics. Furthermore, there is no evidence that such a method produces a more accurate picture of translation quality than the more subjective method used in the present study. We also note that scoring sentences is different from scoring paragraphs or larger passages, and therefore, many previously used evaluation methods are not appropriate.

Finally, we did not compare these results with those resulting from automated evaluations such as BLEU (Papineni et al. 2001). Initially, the focus of our rating system was not to compare two translation systems, or to focus on word-matching metrics, but to see if basic, communicative quality of sentences could be measured effectively. Going forward, however, we will compare automated rankings of sentences from our larger corpus to human rankings. We welcome further research into MT evaluation methods specifically for rating sentences, in addition to further investigations into the effect of scoring methods on tester agreement results.

The research reported in this document/presentation was performed in connection with Contract Number DAAD17-01-C-0069 with the U.S. Army Research Laboratory. The views and conclusions contained in this document/presentation are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of the U.S. Army Research Laboratory or the U.S. Government unless so designated by other authorized documents. Citation of manufacturer's or trade names does

[12] Since precision and recall (upon which Brants' F-score, the harmonic mean was based) cannot properly be calculated for our small sample, we can only compare agreement rates here, not F-scores.

[13] Shows mean scores for combined categories

not constitute an official endorsement or approval of the use thereof. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

Y. Akiba, K. Imamura, K, and Eiichiro Sumita. 2001. Using Multiple Edit Distances to Automatically Rank Machine Translation Output, In *Proceedings of MT Summit VIII, Santiago de Compostela, Spain*

L. Ahrenberg and M.Merkel. 2000. Correspondence measures for MT evaluation. In *Proceedings of the LREC 2000 Workshop on Evaluation of Machine Translation,* Athens, Greece 29th May, 2000, pp. 41-46.

T. Brants. 2000. Inter-Annotator Agreement for a German Newspaper Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation* LREC, 2000, Athens, Greece

J. Carletta 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 222 249-254

Keith Miller and Michelle Vanni. 2001. Scaling the ISLE taxonomy: development of metrics for the multi-dimensional characterization of machine-translation quality. In Proceedings of the Workshop on Example-based Machine Translation, MT Summit VIII, Santiago, Spain

E. Nyberg 3rd , T. Mitamura and J. Carbonell. 1994. Evaluation Metrics for Knowledge-based Machine Translation. In *Proceedings of COLING-94*, Kyoto, Japan

R., J. Passonneau and Diane J. Litman. 1993. Intention-based Segmentation: human reliability and correlation with linguistic cues. I n *Proceedings of the 31st Annual Meeting of the ACL*

K. Papineni, S. Roukos, T. Ward and Wei-Jing Zhu. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation, IBM Report RC22176.

E. Sumita, Setsuo Yamada, Kazuhide Yamamoto, Michael Paul, Hideki Kashioka, Kai Ishikawa, Satoshi Shirai. 1999. Solutions to Problems Inherent in Spoken-language Translation: The ATR-MATRIX Approach., *In Proceedings of the Machine Translation Summit VII, pp. 229-235, Singapore.*

G. Van Slype. 1979. Critical Study of Methods for Evaluating the Quality of Machine Translation. Final Report, Bureau Marcel van Dijk / European Commission, Brussels.

Jean Veronis 1999. A study of polysemy judgments and inter-annotator agreement, In *Programme and advanced papers of the Senseval Workshop.* Herstmonceux Castle, England

A. Voutilainen. 1999. An Experiment on the Upper Bound of Inter-judge Agreement: the case of tagging., In *Proceedings of the European Association of Computational Linguistics 1999*